DOCUMENT RESUME

ED 050 140 TM 000 527

AUTHOR Branson, Robert K.
TITLE Formative Evaluation Procedures Used in Designing a Multi-Media Physics Course.
INSTITUTION Florida State Univ., Tallahassee.
PUB DATE Oct 71
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association, New York, New York, February 1971

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS Audiovisual Aids, College Students, *Course Objectives, Criterion Referenced Tests, *Formative Evaluation, Instructional Materials, *Multimedia Instruction, *Physics, *Program Development, Programed Instruction, Teaching Methods
IDENTIFIERS *United States Naval Academy

ABSTRACT

This research conducted at the U.S. Naval Academy as part of a total effort to design a multi-media physics course, collected data on the learning materials used in Fall 1969, investigated the technical characteristics of the criterion-referenced test items used, and studied student preferences for the alternative study approaches. The ultimate intention is to develop an effective and efficient physics course which can be readily modified by the various course instructors. Seven alternative teaching/learning approaches were developed, based on objectives derived from four widely used physics textbooks. These approaches included two forms of lectures, a study guide, videotapes, and other audiovisual aids. Results demonstrated that the method of instruction was not the critical element in student performance, that students could achieve good results on their own if provided with the necessary instruction and materials, and that if data is collected systematically and used to revise course components, considerable improvements can be made at each iteration. (CK)

C 101

FORMATIVE EVALUATION PROCEDURES USED IN DESIGNING A MULTI-MEDIA PHYSICS COURSE[1]

Robert K. Branson[2]

Florida State University

Dubin and Taveggia (1966) described an impressive variety of research in which numerous teaching methods have shown virtually no reliable differences in the performance of students. Indeed, if it were not for the finding that study time did contribute to improved grades for college students, their report would have been even more devastating.

Others (e.g. Markle, 1967) have reported dramatic improvements in performance achieved through a systematic approach to course design which concentrates on the learner, not the teacher, which specifies revisions in the materials and approach based on data collected from real students. At least tentatively, it seems reasonable to adopt the position that one may safely ignore the nuances perceived by many as important in "teaching," and concentrate on de-

---

[1] A much more comprehensive set of reports and data have been collected than can be reported here. These data have been selected by the author for the purpose of illustrating specific features of course design. The conclusions drawn are those of the author and do not necessarily reflect the official position of the U. S. Naval Academy, the New York Institute of Technology, or the U. S. Office of Education.

[2] This work was done by the author as a consultant to the New York Institute of Technology in collaboration with Stanley L. Schwartz and William A. Deterline.

signing experiences intended to produce specific performances in learners.

This research was concerned with one aspect of a total effort to design a multi-media physics course, that of collecting data specifically on the learning materials used in the Fall of 1969. Further, it was planned that preliminary comparisions would be made between the traditional physics course and the multi-media course under development.

The 1969 tryout followed two earlier tryouts of similar materials on a much smaller scale. These earlier tryouts were concerned with the level of content and the methodological problems of implementing the multi-media course. The Fall 1969 tryout was designed to collect data on the learning materials and procedures, and the Spring 1970 tryout provided the first opportunity to conduct the course on a relatively self-paced basis.

The original and subsequent versions of the course were designed according to the procedures set forth in the Empirical Course Development Model (Deterline and Branson, 1971), a specific requirement of which is the continuous recycling of the course to insure that planned improvements are made during each iteration. The model provides for the repeated collection and analysis of three distinct types of data: time, performance, and rating.

Time data are collected principally when the subject is actually in contact with the unique feature of a particular experimental condition. Performance data are collected regularly on as saturated a basis as time permits. Rating data are collected on student confidence, preference, and estimates of difficulty. Repeated analyses of this data and continuous revision of the course ultimately should tend to optimize performance according to the requirements found

in subsequent years.

The Fall 1969 tryout was conducted under actual operating conditions at the Academy, except for the rigid scheduling of the students imposed by the experimental design. Ethical considerations and Academy administrative requirements made it essential that no midshipman be penalized because he was a part of the experiment. This rule held true whether the student was a member of the experimental groups or the control group. Thus, the formative data collected during this period are of considerably more value than the summative data.

There are 158 Terminal Objectives (TO's) of the course, each of which is composed of one or more enabling objectives which may or may not be hierarchical. The TO's were ordered according to a strategy developed by Finkel (1969) and learning materials were developed to produce the desired level of student performance on these objectives.

This report is concerned with the specific data collected on each of the learning materials, the technical characteristics of the criterion-referenced test items used as measures of the TO's, and the preferences of the students for the alternative approaches to study. Incidentally, some interim data of a summative nature will be presented to indicate the progress and direction of the continuing work.

The ultimate intention in the project is to provide the Naval Academy with an effective and efficient Physics Course based on the Empirical Development Model, and readily modifiable by the course instructors once the final package has been delivered. It is important to note that the methods and procedures to be employed in

continuously improving the course are thought to be equally important to the final package of materials.

Since it is planned that the course will be revised regularly, any course component should contribute directly, either independently, or in combination, to the achievement of criteria. Other factors being equal, those contributing course components which are least expensive, or, more easily and inexpensively revised, should be selected for the final package. Time, preference, and performance measures should all be considered in assembling the final package.

Hypotheses

Two bases were used to derive hypotheses: The intended purpose of the course, and, the specific characteristics of the parallel media used.

1. The experimental groups will have a higher population mean performance on the Final Examination.

2. The experimental groups will have a smaller population variance than will the control groups on the Final Examination.

3. Audiovisual groups will have a higher population mean score on motion dependent posttest items than will other parrallel media groups pooled.

4. There will be no population mean difference between parallel media groups pooled and other experimental groups pooled on difficult media-related posttest items.

5. There will be no population mean difference between the parallel media groups pooled and the other experimental groups pooled on all media-related posttest items.

Method

SUBJECTS

The subjects were 371 second year midshipmen at the U.S. Naval Academy enrolled in the S211 Physics course during the Fall of 1969. This number includes students who took the Final Examination and on whom background data was available.

MATERIALS

The learning materials included four widely used basic Physics textbooks selected by the Faculty of the U.S. Naval Academy for the S211 course. These textbooks served as the source of content from which other learning materials were developed.

Content in the textbooks was selected and converted to Terminal Objectives (TO's). These TO's were then organized and sequenced logically. A Study Guide was developed from the objectives, containing problems and solutions, and additional elaboration of the content, and was programmed with branched remedials according to a scrambled book format.

The content of the Study Guide was then analyzed to select topics for additional elaboration in the "parallel media." The "parallel media" consisted of: a) Videotape recordings of 15-40 minute duration, b) Talking Books, consisting of still visual adaptations of the video portion of the videotape, with an auditory commentary scripted from the audio portion of the videotape, and c) Illustrated Books, in which visuals from the Talking Book were emphasized and highlighted, and, the audio portion was condensed and printed with graphic emphasis.

Four criteria were used to make this selection of content(motion dependent; difficulty; Academy experience; course balance.) This selection procedure is elaborated on in Deterline and Branson (1969).

In addition to the "canned" materials, two types of lecture were

also used. First, a series of lectures with specific demonstrations were prepared and based on the TO's for a given lesson. Then, a series of lectures, without demonstration, was also prepared on the same TO's.

RESPONSE MEASURES

Student performance in the course was measured by 159 criterion referenced constructed response items administered ten items per week except for weeks A, B, and H, where 12, 7, and 20 items respectively were given. Students were asked to make three entries on the answer sheet provided for each question: their answer to the question, their subjective confidence according to a method derived from Shuford and Massengill, and finally by their rating of the item difficulty on a five point scale. Thirty of these 159 items tested TO's on which parallel media presentations were available.

Following the completion of the course, the students were given a 60 item Final Examination composed of half multiple-choice and half constructed-response items. The Final Examination was used principally to determine the student's grade in the course.

The time students spent in contact with the experimental materials or procedures was recorded by proctors. Each Monday, during the second half of the course, students were asked to complete a 13 item rating scale designed to obtain student reactions to specific features of the experimental conditions.

Each test question was rated by the Academy faculty prior to its administration on two five point scales: relative difficulty in Mathematics, and relative difficulty in Physics. Each item was also examined for being inappropriately easy or difficult and for being irrelevent to the TO being tested. An average difficulty rating of Physics and Mathematics was calculated for each test item. Popham and Husek

(1969) have discussed this approach more completely.

The programmed Study Guides used "wet-to-reveal" answer sheets. When the student answered a question, he marked a chemically treated box. If his answer was correct, he was directed to the page containing the next question. If wrong, he was directed to a remedial. When he completed the remedial, he then answered again on the chemical answer sheet. This procedure was repeated until he was correct, or made a total of four responses. The answer sheets were optically scanned and the number of answers actually revealed were counted for each of the 600 plus Study Guide problems.

In addition to the performance, preference, and time data, other measures were available on the midshipmen:

SAT Verbal and Mathematics; Strong Vocational Interest Blank; Quality Point Ratio; Whole Man Score; Physics Department Validating Examination; High School Rank; Academy Naval Scale. It was felt that individual differences might play some part in performance or preference of the various experimental conditions. The rationale for the selection of the measures and the possible implications of their use is detailed in Deterline and Branson (1969), and Branson and Deterline (1971).

PROCEDURE

One-hundred eighty one midshipmen were randomly assigned to seven groups. There were seven experimental conditions allowing each group a different treatment each week for seven weeks. The 16 week Academy semester was divided into two seven-week blocks, allowing for a week of review after each seven week block. The groups were assigned to conditions by the selection of two Latin Squares according to the procedure described in Deterline and Branson (1969). The Latin Squares were used to balance order and sequence effects of presentation, and to permit compliance with Academy policy of making all

materials available to all midshipmen.

Each Saturday, the posttest for that week was given immediately following the administration of the pretest for the coming week. During the semester, the combined testing time for pre and posttests could not exceed one hour.

Each experimental group was identified by the feature distinguishing it as unique: Audiovisual (AV) received the videotapes; Talking Book (TB); Illustrated Book (IB); Lecture Demonstration (LD); Study Guide (SG); Lecture (L); Student Option (SO) was free to use any or none of the prescribed materials.

Posttests were scored immediately upon completion by the students, and scores were posted by 8:00 A. M. Monday. In addition to the midshipman's score, a listing of the TO's on which he missed questions was printed on the same sheet. If a student missed more than three questions, he was requested to make an appointment with his instructor for a remedial session. If he missed more than five, he was required to attend such a session.

One-way analyses of variance were performed on both the confidence modified scores and the proportions correct using, from the Latin Square presentation sequence, groups, weeks, and conditions as treatments. These analyses were performed independently on the media related items and the total posttest scores. Media related items are a subset of the total test score.

Results

HYPOTHESIS TESTS

The null form of Hypothesis 1 was not rejected on the basis of an inadequate t value, $t < 1.0$.

The null form of Hypothesis 2 was rejected on the basis of a variance ratio of 1.58, which with 145 and 188 df, is significant at .01 level.

The null form of Hypothesis 3 was not testable since the sample mean difference was the opposite of the hypothesized direction.

No decision on Hypothesis 4 was made since $\underline{t} < 1.0$.

Hypothesis 5 was rejected (using a two-tailed test) on the basis of a mean difference favoring the non-parallel media conditions ($\underline{t} = -28.36$, 76 df, $p < .005$).

Since the variances were different and the means were not, a possible contributor to that increased variance might be different performance on the two sub-tests contained in the Final Examination. If the distribution of Final Examination scores was bimodal for the control students and not for the experimental students, one possible explanation of the increased variance might be offered. The control group did significantly better on multiple-choice questions than they did on constructed-response questions ($\underline{t} = 4.15$, 382 df, $p < .005$). This difference did not hold for the experimental groups ($\underline{t} = 1.52$, 150 df, $p > .05$).

The one-way analyses of variance were concerned with major group differences on the total scores. First, the F for experimental conditions was not significant, neither for proportion correct nor for confidence modified scores, in both cases being < 1.0.

The F's for weeks were both significant beyond the .01 level, probably indicating, since all treatments occurred in all weeks, a difference in difficulty level of the Physics' materials (proportion correct, $F = 331.5$, df 5, 845, $p < .01$; confidence, $F = 34.6$, df 5, 845, $p < .01$).

TERMINAL OBJECTIVES

For each of the TO's, as represented by the posttest items and the Study Guide responses, the following data were collected for re-

vision purposes:

proportion correct

confidence modified scores

difficulty rating in Physics

difficulty rating in Mathematics

student difficulty rating

recorded confidence

Insert Table 1 about here

Table 1 indicates the correlation matrix among these variables. Faculty Physics and Mathematics ratings show moderately high (.66) linear relationship. Also of note is the r of .66 between the student's confidence and his rating of difficulty. He was more confident of the difficult items. It also appears that the faculty was a better judge of the difficulty of the items than was the student. All of these relationships are high enough to be useful in course revision.

BACKGROUND VARIABLES

In searching for data upon which to base media decisions, it was thought that an analysis of the relationships among background and performance variables would improve predictions. Table 2 indicates the intercorrelations among 9 selected variables. While none of the correlations is surprisingly high or low,

---

Insert Table 2 about here

---

the lack of relationship between the Final Examination and the total posttest performance is encouraging, in light of the intentions of the course designers.

PREFERENCE and TIME DATA

Items on the rating scale were combined for each experimental

group and a single dimension of favorable-neutral-unfavorable was used for between group comparisons. Each group could be described with three scores: proportion favorable, proportion neutral, and proportion unfavorable. Two rankings were made, the first ranked the groups on the favorable proportion, the second ranked the groups on the unfavorable proportion.

Insert Table 3 about here

Table 3 reveals the Lecture and the Student Option conditions to be essentially tied and rated both as most favorable and least unfavorable. The L/SG condition earned the opposite rankings: least favorable and most unfavorable.

While the favorable and unfavorable are two ends of the same continum, and are thus not independent, both rankings were made to eliminate the influence of neutral responses.

Total time spent in each experimental condition by all students was computed and rank-ordered from least to most. Mean time per condition was 171 minutes with a standard deviation of 61 minutes. Lecture and Student Option conditions required the least amount of time, while the L/SG condition required the most. The rank correlation (rho) between preference and time was .67, allowing rejection of the null hypothesis (Ho: rho=0) at the .02 level.

Discussion

PERFORMANCE DATA

It was not the purpose of the tryout to arrange experimental conditions which would produce statistically significant differences. Rather, the purpose was to gather data which would be useful in revising the course to make it more appropriate for student-paced use. Procedurally, the kinds of data collected must be relatively inexpen-

sive and require minimum time. Successive iterations of the course are not likely to improve student performance if such data cannot be used for purposes of revision.

It is one purpose of the course to increase mean student performance and reduce the variation in group performance. To that end, the rejection of Hypothesis 4 indicates that progress was made. While the difference between means favored the experimental groups, the difference was not statistically significant.

Not unexpectedly, all experimental groups did as well on those Final Examination questions requiring constructed response answers as they did on the multiple-choice questions. This was not true of the control groups, even though no correction for guessing was made.

The experimental conditions were all apparently equally effective in teaching students the required criterion behavior. It should be noted, however, that the criteria were based on a highly limited range of responses: the working of Physics problems. This conclusion seems warranted, regardless of whether one uses the norm-referenced Final Examination, or the criterion referenced total posttest scores. In the special case of the media related test items, the non-audiovisual groups did significantly better in total performance.

If one considers the performance data in light of the preference data, it appears that students are concerned with those experimental conditions which take the least time and which are most directly related to the content of the tests. For example, the L/SG condition was considerably less attractive to the students than was the straight Lecture (L) group. Conceivably, while the demonstration may have been interesting, the students viewed it as having no relationship to the important criteria of the course, namely, the working of Physics pro-

blems. While the inadequacy of such criteria has been discussed more fully elsewhere (Branson, 1970) they are, nevertheless, widely used.

The preference of the SO condition may be attributable to the small amount of time actually prescribed for the students during those weeks. That is, if the lecturer is willing to show students how to work problems, he is willing to listen. However, if one burdens the student with demonstrations, Audiovisual presentations, etc., the student seems much more willing to do it himself.

Regardless of the intergroup comparisons, the data collected are quite interesting. Each Terminal Objective was treated in a variety of ways: in the Study Guides, textbooks, and the lectures. The criterion referenced test items used to measure the behavior were evaluated by the faculty along a number of dimensions: appropriateness to the TO (content validity), difficulty in Mathematics, difficulty in Physics.

These ratings are extremely valuable in providing a methodology by which a faculty member can, a priori, determine the level at which his course is taught. Provided that one is willing to accept final performance of the students as an indication of the level of sophistication of the course, the degree to which this can be predicted in advance is a good indicator of the course "level."

If, on the other hand, it is necessary to wait until after the results are in to specify the level, it appears that the students, not the faculty, decide what level of performance is acceptable. Particularly, if the grades in the course are assigned on any "normal" curve basis.

Our results indicate that the faculty is considerably better at predicting student performance on the basis of difficulty ratings than

the students are. Faculty correlations were -.42 and -.58 between performance and difficulty, while students difficulty and performance correlated only -.25.

This procedure for establishing course difficulty level appears emminently more desirable than a method which uses ad hoc student performance, to determine which test items should be retained and discarded. Our results indicated that there was a significant "weeks" effect, from which we inferred that weeks were not equally difficult. Physicists confronted with this data claimed to have known all along that some topics were indeed more difficult than others, as is virtually always the case in academic subjects.

The fact that they could predict, with reasonable precision, the level of difficulty of the test items, and, thus, control this level of difficulty, transfers the responsibility of course level determination to the faculty..

The Study Guide results were of great general interest. While the "Linear-Branching" programmed instruction controversy has been dead for many years, it appeared reasonable in this course to offer specific remedial frames, to which the student was looped, when he failed to answer correctly on the first attempt. Further, that more specific remedials would be more effective than general remedials. While the data for each Volume of the Study Guide has been presented elsewhere, an analysis of Volume N is interesting at this point. Volume N had "general" remedials. That is, the remedial was simply a presentation of the correct way to work the problem. The remainder of the course used specific remedials. That is, each problem was analyzed and the most likely, common, and probable errors were selected for elaboration. The students were shown why they were wrong, not

how to do the problem correctly.

If a remedial is effective, it ought to reduce the probability of error on the subsequent attempts at the answer. Thus, if a student has missed the correct answer on the first trial and is given a remedial, he ought to have a better chance to be right on the second attempt than someone not receiving the specific remedial. We have used the following formula to assess the effectiveness of remedials, where

$$\text{Effectiveness} = \frac{\text{number of double choices}}{\text{total number of double, triple, and quadruple choices}}$$

The results for Volume N indicated an effectiveness index of .59 and the general results of the course indicated an effectiveness index for the remaining Volumes of .60. On the basis of this data, it was decided not to include specific remedials dealing with student errors in subsequent versions of the course. Course developers would concentrate on a more careful description of the correct way of working the problems.

Finally, the very low correlation between the performance of students on the total of 159 criterion referenced items and the 60 item norm-referenced Final is encouraging. Professors judgment of performance on criterion referenced items is a better indicator of final score on these items than is total student performance on norm-referenced items used as a predicator. Since the posttest items had been carefully screened for content validity prior to their inclusion on the test, and had been judged according to their expected level of difficulty, it was possible to make a more accurate determination of the actual course level of difficulty than would otherwise have been possible.

Subsequent versions of the course can use the test items in a pretest form and establish a baseline of student performance, having available past performance on the same items as a comparison. It is important to note here that professor judgment, tempered by past experience, is the critical element in developing the criterion measures. Student performance alone is not used. Consequently, test items are not discarded when a large number or proportion of students answers them correctly. They are discarded when they are rated and judged inappropriate by the faculty.

The results of the Fall 1969 tryout demonstrated to the Physics' faculty that the method of instruction was not the critical element in student performance, an accomplishment of some magnitude. Further, that students could, when provided with the necessary instruction and materials, achieve good results on their own.. And finally, that if data is collected systematically and used to revise the course components, improvements can be made at each successive iteration..

References

Branson, R. K. The criterion problem in programmed instruction. Educational Technology, 1970, X(7), 35-37.

Branson, R. K. & Deterline, W.A. Final report: multi-media physics course. (Technical Report No. 5.0) Old Westbury, New York,1971. New York Institute of Technology, 197 , in preparation.

Deterline, W. A. & Branson, R. K. Evaluation and validation design. (Technical Report No. 4.7) Old Westbury, New York , New York Institute of Technology, 1969.

Deterline, W. A. & Branson, R. K. An empirical course development model. (Technical Report No. 5.7) Old Westbury, New York, 1971. New York Institute of Technology, 197

Dubin, R. & Taveggia, T. C. The teaching-learning paradox. Eugene, Oregon: Center for the Advanced Study of Educational Administration, 1968.

Finkel, R. Rationale for sequencing objectives. (Technical Report No. 3.5) Old Westbury, New York , New York Institute of Technology, 1969.

Markle, D. G. Final report: the development of the bell system first aid and personal safety course. Palo Alto, California: American Institutes for Research, 1967.

Popham, W. J. & Husek, T. R. Implications of criterion referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Table 1

Correlation Matrix for the Measures

Taken on Each of the Posttest

Items

Variable	1	2	3	4	5
	Performance Mean Log Confidence	Math Difficulty	Physics Difficulty	Mean Confidence	Student Difficulty
1		-.42**	-.58**	.03	-.25**
2			.66**	-.11	.10
3				.07	.24**
4					.66**
5					

Note.-

** $p < .01$

The negative correlations indicate that as difficulty ratings increase, performance decreases.

Table 2

Intercorrelations of Background and Performation Variables on Those Subjects From Whom a Complete Set of Data Was Available N = 77

	1	2	3	4	5	6	7	8
1. SAT Verbal								
2. SAT Math	.31							
3. Highschool Rank	.14	.06						
4. Whole Man	.17	.32	.57					
5. Quality Point Ratio	.22	.28	.43	.53				
6. Final Exam	.34	.39	.25	.27.	.70			
7. Physics Validation	.23	.36	.20	.23.	.38	.52		
8. Media Related	.05	.12	.03	.22	.28	.23	.04	
9. Total Posttest	-.03	.08	.09	.35	.40	.25	.11	.74

Note.- For 70 df, the .05 level is .23, the .01 level is .30.

Table 3

Preferences and Time Data and the Rank Orderings of Their Proportions

	Proportions			Rank Order		Time
	Favorable	Neutral	Unfavorable	Most Favorable	Least Unfavorable	(Least To Most)
Lecture	.50*	.32*	.18*	1	1	2
Student Option	.47*	.34*	.19*	2	2	1
Talking Book	.42	.35	.23	3	3	4
Study Guide	.41	.35	.24	4	4	3
Audio-Visual	.34	.41	.25	5	5	6
Illustrated Book	.33	.40	.27	6	6	5
Lecture/ Demonstration	.26*	.40*	.34*	7	7	7
Mean	.39	.37	.24			
Standard Deviation	.08	.03	.05			

Note.-

* Indicates a deviation of ±1 S.D.